Certain Query Answering on Compressed String Patterns: From Streams to Hyperstreams (long version)

Iovka Boneva, Joachim Niehren, Momar Sakho

Certain Query Answering on Compressed String Patterns: From Streams to Hyperstreams

Iovka Boneva[1], Joachim Niehren[2], and Momar Sakho[2]

[1] Université de Lille, France
[2] Inria Lille, France

Abstract. We study the problem of certain query answering (CQA) on compressed string patterns. These are incomplete singleton context-free grammars, that can model systems of multiple streams with references to others, called hyperstreams more recently. In order to capture regular path queries on strings, we consider nondeterministic finite automata (NFAs) for query definition. It turns out that CQA for Boolean NFA queries is equivalent to regular string pattern inclusion, i.e., whether all strings completing a compressed string pattern belong to a regular language. We prove that CQA on compressed string patterns is PSPACE-complete for NFA queries. The PSPACE-hardness even applies to Boolean queries defined by deterministic finite automata (DFAs) and without compression. We also show that CQA on compressed linear string patterns can be solved in PTIME for DFA queries.

1 Introduction

A stream is a sequence of events that arrive incrementally one by one from the left to the right. Most typically, streams are produced by social networks such as Twitter, database systems as for producing financial transactions, information systems, sensor systems, or more generally when communicating semi-structured data over the internet. We are interested in the problem of monitoring streams in a reactive manner [22,16,25,23]. The objective is to select the relevant events of a stream as quickly as possible upon their arrival. This requires to decide whether an event of the stream is a certain answer of the logical query that defines the relevant events of the monitoring task. Lowering the latency of this decision process increases the reactivity of the stream processing system and reduces its memory costs. A limitation to constant memory may seem ideal in theory, but is too restrictive for many monitoring tasks in practice. A less restrictive objective is thus to minimize the latency and the memory consumption.

In the present paper we study a generalization of streams to multiple streams with references as introduced by Maneth, Pereira and Seidl [21]. The references point to unknown parts in the middle of a stream. The same reference may be used multiple times, allowing to share unknown parts. Streams with similar references were named hyperstreams in own previous work [20]. Here, we propose to formalize hyperstream containing words (rather than linearizations of trees or nested words) as *incomplete* versions of singleton context-free grammars [24]

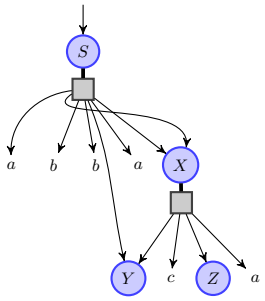

Fig. 1: Hyperstream G_1.

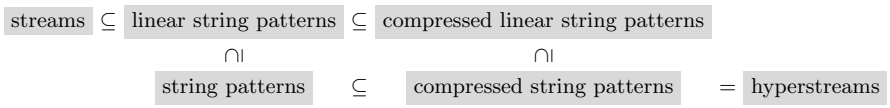

Fig. 2: Landscape from streams to hyperstreams.

(also termed straight line programs [3]), where the rules of some nonterminals may be missing. The hyperstream G_1 illustrated graphically in Fig. 1 has the terminals in $\{a, b, c\}$, the nonterminals in $\{S, X, Y, Z\}$, and the rules $S \to aXbbYaX$ and $X \to YcZa$. The nonterminals are called the references of the hyperstream. For some of these references there exists a rule in the grammar, and if so, this rule is unique. For any grammar rule, the reference on its left is said to refer to the string pattern on its right. The hyperstream G_1 has a rule for S and X, while it misses those of Y and Z. The missing rules for these references may be added in the future one by one by the hyperstream's environment.

Alternatively, hyperstreams can be identified with *compressed string patterns*. The hyperstream G_1 for instance represents the string pattern

$$p_1 = a\underline{YcZa}bbYa\underline{YcZa},$$

while sharing the underlined factors substituted for the two occurrences of X. Streams are a special case of string patterns that have a unique occurrence of a variable in their last position. The landscape from streams to hyperstreams, over linear string patterns, string patterns, and compressed string patterns is illustrated in Fig. 2.

In this paper, we study the decision problem of certain query answering (CQA) on compressed string patterns, i.e., whether a tuple of positions is a certain answer of a query on a compressed string pattern. Here we consider the positions of the string pattern after decompression rather than the positions of the grammar. Intuitively, a tuple of positions is a certain query answer on a compressed string pattern G if it is an answer to the query on all completions of G, up to the offsets raised by the completion of G on its decompression. We will also consider the symmetric problem for certain query non-answers.

Motivated by regular path queries [1], we consider nondeterministic finite automata (NFAs) for defining such queries. For instance, the query $\mathbf{Q}_1$ on strings over $\Sigma = \{a, b, c\}$ that selects all a-positions that are followed eventually by bb can be defined by the following regular path in XPATH-like notation:

$$\mathbf{Q}_1 = \texttt{successor}^*\texttt{::}a[\texttt{successor}^*\texttt{::}b/\texttt{successor::}b].$$

It can also be defined by the x-pointed regular expression $\Sigma^* a^x \Sigma^* bb\Sigma^*$ where x is a variable for the position that is to be selected. Now, consider the case, where the string is not given explicitly but only described partially by some (compressed) string pattern. On the string pattern p_1, for instance, the a-positions 1 and 5 are certain query answers for $\mathbf{Q}_1$, while the a-positions 9 and 13 are not. The position 13 is even a certain non-answer.

When restricted to Boolean NFA queries, CQA becomes equivalent to the problem of whether all strings described by the completions of a compressed string pattern are accepted by the NFA. For the string pattern Y (for some variable Y), this problem clearly generalizes on the universality problem of NFAs, which is well-known to be PSPACE-complete [17]. The following questions, however, are open to the best of our knowledge, even in the case of string patterns without compression: Is CQA on (compressed) string patterns decidable for NFA-defined queries, and if yes, what is the complexity? Does CQA on (compressed) string patterns remain hard for queries defined by deterministic finite automata (DFAs)? For which restrictions of (compressed) string patterns is CQA in PTIME? And what about the symmetric questions concerning certain query non-answers? The objective of the present paper is to answer these questions in all possible cases.

Our first contribution is that CQA on string patterns is PSPACE-complete, both for NFA queries and DFA queries, with and without compression, Boolean or not, see Fig. 3. This upper bound is not fully obvious, as the set of strings defined by a string pattern may be non-regular and even non-context-free. Furthermore, the lower bound may be surprising in that CQA for DFA queries on string patterns is more complex than on streams, where it is in PTIME (Theorem 1 of [13]), and also more complex than string pattern matching, which is NP-complete (Theorem 3.6 of [2]) even with compression (Theorem 4.10 of [11]).

Our second contribution is that CQA for DFA queries can be decided in PTIME on compressed *linear* string patterns, see Fig. 4. The linearity restriction matches with the worst case complexity for streams, even though linear compressed string patterns allow for unknown factors and compression in addition. This result (Corollary 2) is based on a novel algorithm for partial decompression of compressed string patterns that we present (Lemma 6), followed by a test of a reachability property (Theorem 3).

Our third contribution is that the certainty of query non-answers on compressed string patterns is PSPACE-complete, both for Boolean and non-Boolean queries, and independently of whether they are defined by DFAs or NFAs. In the Boolean case, the problem is equivalent to whether a compressed string pattern does *not* match the regular language accepted by the automaton. This problem

	DFAs	NFAs
Answers	PSPACE-c	PSPACE-c
Non-answers	PSPACE-c	PSPACE-c

Fig. 3: Query certainty on (compressed) string patterns or hyperstreams.

	DFAs	NFAs
Answers	PTIME	PSPACE-c
Non-answers	PTIME	PTIME

Fig. 4: Query certainty on (compressed) linear string patterns or streams.

generalizes on the complement of compressed string pattern matching, and thus is coNP-hard. So while certain query non-answering can be solved in PTIME on streams, the complexity increases to PSPACE on compressed string patterns. Finally, we show that the restriction of the problem to compressed linear string patterns – that is, *regular* compressed linear string pattern matching – can also be solved in PTIME even for queries defined by NFAs.

Outline. In Section 2 we start with some preliminaries on finite automata theory. Section 3 recalls the notion of compressed string patterns and in Section 4 we study the problems of regular compressed pattern inclusion and matching. Section 5 recalls how to define non-Boolean queries on strings by automata. In Sections 6 and 7 we generalize the notions of certain query answers and non-answers to (compressed) string patterns and study their complexity.

Related Work. The notion of certain query answers for incomplete relational structures is standard in databases research [9]. In the context of stream processing, certain query answers were called answers that are safe for selection and certain query non-answers were called safe for rejection [12]. Certain query non-answers were studied for fast failure [4] and for reducing the memory consumption of streaming systems. The problem of certain query answering and non-answering on streams has been shown to be computationally hard even for queries defined in tiny fragments of first-order logic [12]. Certain query non-answering was shown to be hard in the context of online verification [4,19].

As shown by [12], those classes of queries on strings for which the problem of certain query answering on streams is known to be feasible, are either such that certainty is always determined with 0-delay [22,4,14]) or such that the queries in the class can be compiled to DFAs in PTIME.

Algorithms for processing XML streams or complex event streams raised much interest in the literature [16,25,15] and motivated the work on hyperstreams. XML streams contain nested words [22,18] rather than strings without bracket structure. The best existing algorithms for answering navigational XPATH queries (i.e. first-order logic queries) on XML streams are based on compilation to nested word automata [10,23]. Low but not lowest latency is achieved with high efficiency by approximating certain answers for queries defined by nondeterministic nested word automata.

2 Preliminaries

Words. The set of natural numbers with 0 is denoted by $\mathbb{N}$. For any set Σ, a word over Σ is a tuple $(a_1, \ldots, a_n) \in \Sigma^n$ where $n \in \mathbb{N}$. We denote such words

by $a_1 \dots a_n$ and by ε if $n = 0$. We denote the i^{th} letter of a word $u = a_1 ... a_n$ by $u[i] =_{def} a_i$. The set of all words over Σ is denoted by Σ^*. The concatenation of two words $u_1, u_2 \in \Sigma^*$ is denoted by $u_1 \cdot u_2 \in \Sigma^*$. For instance, if $\Sigma = \{a, b\}$ then $aba \cdot a = abaa$. The set of positions of a word $u = a_1 \dots a_n$ is $pos(u) = \{1, \dots, n\}$. For any subset $\Sigma' \subseteq \Sigma$ the set $pos_{\Sigma'}(u)$ is the subset of positions i of u such that $a_i \in \Sigma'$. Given a word $w = a_1 \dots a_n$ and a second word $u = b_1 \dots b_n$ of the same length possibly on a different alphabet, we define the zipped word over the product alphabet by $w * u = (a_1, b_1) \dots (a_n, b_n)$. As a convention throughout the paper, we use the term *string* for words over the default alphabet Σ, as opposed to words over other alphabets such as $\Sigma \cup \mathcal{Y}$ for string patterns, and $\Sigma^{\mathcal{V}}$ for $\mathcal{V}$-annotated strings, that will be introduced later on.

Monoids. A monoid is a triple $(M, \cdot^M, 1^M)$ where M is a set, $\cdot^M : M \times M \to M$ is an associative binary operation and $1^M \in M$ is the neutral element, i.e. $1^M \cdot^M m = m \cdot^M 1^M = m$ for all $m \in M$. Given a word $u = m_1 \dots m_n \in M^*$ we define its evaluation by $u^M = m_1 \cdot^M \dots \cdot^M m_n$ where $\varepsilon^M = 1^M$.

Most typically, we will consider the monoid of words $(\Sigma^*, \cdot, \varepsilon)$ on some alphabet Σ, with the concatenation operation $\cdot : \Sigma^* \times \Sigma^* \to \Sigma^*$, and the empty word ε as neutral element. Alternatively, given another set Q, we will consider the *transition monoid* $(T_Q, \circ, id)$, where $T_Q = 2^{Q \times Q}$ is the set of binary relations over Q, $\circ : T_Q \times T_Q \to T_Q$ is the composition operation of binary relations on Q, and $id = \{(q, q) \mid q \in Q\}$ is the identity transition.

Finite Automata. A nondeterministic finite-state automaton (NFA) is a tuple $A = (Q, \Sigma, \delta, I, F)$ where Q and Σ are finite sets, $\delta \subseteq Q \times \Sigma \times Q$, and $I, F \subseteq Q$. We call Q the state set, Σ the alphabet, δ the transition relation, I the set of initial states, and F the set of final states of the automaton. An NFA A is deterministic, or a DFA, if it has exactly one initial state and the transition relation δ forms a partial function from $Q \times \Sigma$ to Q. The elements of T_Q are called the *transitions* of A. Any transition relation $\delta : Q \times \Sigma \times Q$ can be extended homomorphically to a transition function $\delta : \Sigma^* \to T_Q$ that assigns to any string a transition of A. Here we overload the symbol δ to stand for the transition relation and the transition function. The transition $\delta(a)$ of a letter $a \in \Sigma$ is $\{(q, q') \mid (q, a, q') \in \delta\}$ and the transition $\delta(w)$ of a string $w = a_1 \dots a_n \in \Sigma^*$ is $\delta(w) = (\delta(a_1) \dots \delta(a_n))^{T_Q}$. This is the composition of the transitions of all letters of w based on the operations of the transition monoid T_Q, and its neutral element for the empty word. Note that if A is a DFA then all transitions $\delta(w)$ are partial functions. A transition τ is called *I,F-successful* if $\tau \cap (I \times F) \neq \emptyset$. The language of an NFA $A = (Q, \Sigma, \delta, I, F)$ is the set $\mathcal{L}(A) = \{w \in \Sigma^* \mid \delta(w)$ is I,F-successful$\}$. The size of an automaton is $|A| = |Q| + |\delta|$.

Example 1. NFAs can be used to define Boolean queries on strings such as query $\mathbf{Q}_2$ which tests whether some position will be selected by query $\mathbf{Q}_1$, i.e., whether a string contains an a-letter followed eventually by a factor bb. The language of all such strings is defined by the automaton on Fig. 5.

A transition τ is called *δ-inhabited* if there exists a word $w \in \Sigma^*$ such that $\delta(w) = \tau$. *Transition inhabitation* Inh_Σ is the decision problem that receives

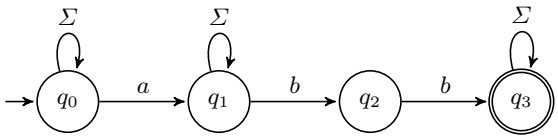

Fig. 5: Automaton A_2 defining the Boolean query $\mathbf{Q}_2$ with alphabet $\Sigma = \{a, b, c\}$

as input a finite set Q, a transition relation $\delta \subseteq Q \times \Sigma \times Q$ and a transition $\tau \in T_Q$, and outputs whether τ is δ-inhabited. Inh_Σ is also called the *membership problem of the transition monoid* $\delta(\Sigma^*) \subseteq T_Q$ [7].

Theorem 1 (Kozen [17]). *For any set Σ with at least 2 elements, the transition inhabitation problem Inh_Σ is* PSPACE*-complete.*

The PSPACE hardness proof can be done by reduction from the problem of *non-emptiness of the intersection of sequences of* DFA*s*, which was shown to be PSPACE-complete in [17] too.

3 Compressed String Patterns

We fix an infinite set $\mathcal{Y}$ of *string variables* for the rest of the paper. A *string pattern* over a finite alphabet Σ is a word in $(\Sigma \cup \mathcal{Y})^*$. The set of all string patterns over Σ is denoted by Pat_Σ. The set of variables that occur in a string pattern p is denoted by $fv(p)$. An instance of a string pattern $p \in Pat_\Sigma$ is a string that can be obtained by substituting the variables of p by strings in Σ^*. Any substitution $\sigma : \mathcal{Y} \to \Sigma^*$ can be lifted to a substitution on string patterns $\hat{\sigma} : Pat_\Sigma \to \Sigma^*$ such that for all $p, p' \in Pat_\Sigma$, $a \in \Sigma$, and $Y \in \mathcal{Y}$: $\hat{\sigma}(pp') = \hat{\sigma}(p) \cdot \hat{\sigma}(p')$, $\hat{\sigma}(\varepsilon) = \varepsilon$, $\hat{\sigma}(a) = a$, and $\hat{\sigma}(Y) = \sigma(Y)$. We define the set of *instances* of a string pattern $p \in Pat_\Sigma$ as:

$$Inst(p) = \{\hat{\sigma}(p) \mid \sigma : \mathcal{Y} \to \Sigma^*\}.$$

For example, the string $a\underline{cb}\underline{cb}abb\underline{cb}a$ is an instance of the pattern $aYcZabbYa$, obtained with the substitution $[Y/cb, Z/b]$. A string pattern is called *linear*, if all its variables occur at most once. The set of all linear string patterns over Σ is denoted $LinPat_\Sigma$.

Definition 1 (Compressed string pattern). *A compressed string pattern is an* acyclic *CFG $G = (N, \Sigma, R, S)$ where $N \subseteq \mathcal{Y}$ is a finite set of nonterminals, Σ is an alphabet of terminal symbols disjoint from $\mathcal{Y}$, the ruling function R is a partial function that maps some of the nonterminals in N to string patterns in $(N \cup \Sigma)^*$, and $S \in N$ is the start symbol. The set of all compressed string patterns over Σ is denoted by $cPat_\Sigma$.*

We recall that a CFG G is acyclic if the binary relation $>_G = \{(Y, Z) \mid Y \in dom(R), Z \in fv(R(Y))\}$ is acyclic. The compressed string pattern from the introduction for instance has the rules $R(S) = aXbbYaX, R(X) = YcZa$. These rules induce the binary relation $\{(S, X), (S, Y), (X, Y), (X, Z)\}$ which is acyclic. The size of G is $|G| = |N| + \sum_{Y \in dom(R)} |R(Y)|$. The set of free variables of G is $fv(G) = N \setminus dom(R)$. A compressed string pattern G is called a *singleton context-free grammar* (sCFG) if it has no free variables, that is $fv(G) = \emptyset$. It is well-known that any sCFG defines a single string in Σ^*. The object of interest here is the set of strings that can be obtained by completing a compressed string pattern G to a sCFG, or equivalently, the set of instances of the string pattern of G defined as follows. For any compressed string pattern $G = (N, \Sigma, R, S)$, the grammar $G' = (N \setminus fv(G), \Sigma \cup fv(G), R, S)$ is a sCFG. We define the string pattern $pat(G) \in Pat_{\Sigma \cup fv(G)}$ as the unique word in the language of G'. Formally, for every $Y \in dom(R)$, let G_Y be the compressed string pattern $G_Y = (N, \Sigma, R, Y)$. If $S \in dom(R)$ then $pat(G) = \hat{\sigma}(R(S))$ where $\sigma(Y) = pat(G_Y)$ for all $Y \in dom(R) \setminus \{S\}$. This recursive definition is well-founded because G is an acyclic CFG. Otherwise, if $S \in fv(G)$, then $pat(G) = S$. For instance, if G_1 is the hyperstream from the introduction, then $pat(G_1) = p_1$.

A compressed string pattern G is called a *compressed linear string pattern* if $pat(G)$ is linear. The set of all compressed linear string patterns over Σ is denoted $cLinPat_\Sigma$. Finally for any string pattern $p \in Pat_\Sigma$ there exists a compressed string pattern G_p having p as pattern, namely $G_p = (\{S\} \cup fv(p), \Sigma, R, S)$ with $dom(R) = \{S\}$ and $R(S) = p$. Clearly $pat(G_p) = p$. Therefore we will identify p with G_p, so that $Pat_\Sigma \subseteq cPat_\Sigma$.

4 Regular Pattern Inclusion and Matching

We consider the problems of *regular compressed pattern inclusion*, i.e. testing whether all strings described by a completion of a compressed string pattern to a sCFG are accepted by a finite automaton, and of *regular compressed pattern matching*, whether some string described by a completion of a compressed string pattern is accepted by a finite automaton.

A class of compressed string patterns $\mathcal{G}$ is a function from finite sets Σ to subsets $\mathcal{G}_\Sigma \subseteq cPat_\Sigma$ such as for instance $\mathcal{G} \in \{Pat, cPat, LinPat, cLinPat\}$. A class of NFAs $\mathcal{A}$ is a function from finite sets Σ to subsets $\mathcal{A}_\Sigma \subseteq \text{NFA}_\Sigma$, where NFA_Σ is the set of NFAs with alphabet Σ. Most typically, $\mathcal{A} \in \{\text{DFA}, \text{NFA}\}$. For any class $\mathcal{G}$ of compressed string patterns, any class $\mathcal{A}$ of NFAs, and any finite set Σ we define the following two decision problems.

Regular compressed pattern inclusion $\text{INCL}_\Sigma(\mathcal{G}, \mathcal{A})$. *Input:* A compressed string pattern $G \in \mathcal{G}_\Sigma$ and a finite automaton $A \in \mathcal{A}_\Sigma$.
Output: The truth value of whether $Inst(pat(G)) \subseteq \mathcal{L}(A)$.

Regular compressed pattern matching $\text{MATCH}_\Sigma(\mathcal{G}, \mathcal{A})$. *Input:* A compressed string pattern $G \in \mathcal{G}_\Sigma$ and a finite automaton $A \in \mathcal{A}_\Sigma$.
Output: The truth value of whether $Inst(pat(G)) \cap \mathcal{L}(A) \neq \emptyset$.

The problem $\text{coMATCH}_\Sigma(\mathcal{G}, \mathcal{A})$ is the complement of the problem $\text{MATCH}_\Sigma(\mathcal{G}, \mathcal{A})$, and thus outputs for a given compressed string pattern $G \in \mathcal{G}_\Sigma$ and a finite automaton $A \in \mathcal{A}_\Sigma$ whether $Inst(pat(G)) \cap \mathcal{L}(A) = \emptyset$.

Example 2. Any instance of $pat(G_1) = aYcZabbYaYcZa$ answers the Boolean query $\mathbf{Q}_2 = [\texttt{successor}^*\texttt{::}a/\texttt{successor}^*\texttt{::}b/\texttt{successor::}b]$ from Example 1, i.e., the instance set of $pat(G_1)$ is included in the language of NFA A_2 in Fig. 5.

Let sDFA be the subclass of DFAs that recognize a singleton language. Note that the well-known problem of string pattern matching is $\text{MATCH}_\Sigma(Pat, s\text{DFA})$, and $\text{MATCH}_\Sigma(cPat, s\text{DFA})$ is its extension with compression. We recall from [11] that string pattern matching with and without compression respectively are NP-complete for all alphabets Σ with at least 2 letters, but in PTIME when restricted to linear string patterns even with compression.

Our first main contribution is the following complexity result for regular compressed pattern matching and inclusion (see Fig. 3).

Theorem 2 (Non-linear patterns). *For any $\mathcal{G} \in \{Pat, cPat\}$ and $\mathcal{A} \in \{$DFA, NFA$\}$ and for any finite alphabet Σ with at least 2 letters, the problems of regular compressed pattern inclusion $\text{INCL}_\Sigma(\mathcal{G}, \mathcal{A})$ and matching $\text{MATCH}_\Sigma(\mathcal{G}, \mathcal{A})$ are* PSPACE*-complete.*

This shows that these problems are decidable even though the instance sets of nonlinear patterns like $Inst(YYY)$ are neither regular nor context-free. The theorem also shows that regular pattern matching $\text{MATCH}_\Sigma(Pat, \text{DFA})$ is PSPACE-complete and thus harder than compressed string pattern matching $\text{MATCH}_\Sigma(Pat, s\text{DFA})$ which is NP-complete.

Proof. We will present a sequence of PSPACE reductions from Lemma 1 until Lemma 4 that imply the theorem when composed as in Fig. 6.

We introduce the notations $A \leq_p B$ (resp. $A =_p B$) where A and B are decision problems to denote that A reduces polynomially to B (resp. A reduces polynomially to B and B reduces polynomially to A).

Lemma 1. $\text{INCL}_\Sigma(LinPat, \text{NFA})$ *is* PSPACE*-hard if* $|\Sigma| \geq 2$.

Proof. For the linear pattern Y with $Y \in \mathcal{Y}$, the regular compressed pattern inclusion problem $Inst(Y) \subseteq \mathcal{L}(A)$ is equivalent to $\Sigma^* = L(A)$, and universality of NFAs is well-known to be PSPACE-complete if $|\Sigma| \geq 2$.

We now show the PSPACE upper bound for $\text{INCL}_\Sigma(cPat, \text{NFA})$. For any transition relation $\delta \subseteq Q \times \Sigma \times Q$ and any substitution into the transition monoid $\sigma : \mathcal{Y} \to T_Q$, we define δ^σ to be the function that takes a string pattern in Pat_Σ as input, and returns a transition, such that for all $p, p' \in Pat_\Sigma$, $w \in \Sigma^*$, and $Y \in \mathcal{Y}$:

$$\delta^\sigma(w) = \delta(w), \qquad \delta^\sigma(y) = \sigma(y), \qquad \delta^\sigma(pp') = \delta^\sigma(p) \circ \delta^\sigma(p').$$

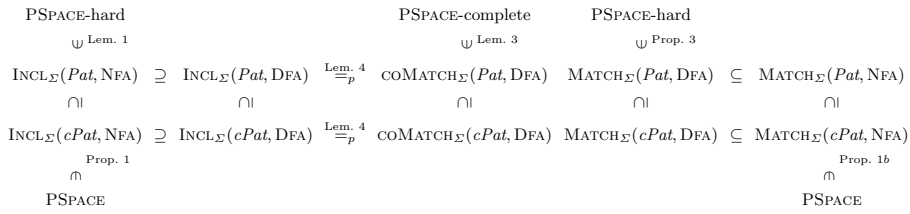

Fig. 6: Regular inclusion and matching problems relationship and complexity classes

Lemma 2. *Given a transition relation $\delta \subseteq Q \times \Sigma \times Q$ of some* NFA*, a substitution $\sigma : \mathcal{Y} \to T_Q$ and a compressed string pattern $G \in cPat_\Sigma$, the transition $\delta^\sigma(pat(G)) \in T_Q$ can be computed in time $O(|Q|^3|G|)$.*

Proof. We represent transitions in T_Q as $|Q| \times |Q|$-matrices. We first precompute the transition matrices $\delta(a)$ for all letters $a \in \Sigma$ in time $O(|\delta| + |\Sigma||Q|^2)$ which is in $O(|Q|^2)$ for fixed Σ. Suppose $G = (N, \Sigma, R, S)$. Let $dom(R) = \{Y_1, \ldots, Y_n\}$ such that if $Y_i >^+_G Y_j$ for some $1 \leq i, j \leq n$ then $i < j$. For all $1 \leq i \leq n+1$ we define:

$$\sigma_{n+1} = \sigma$$
$$\sigma_i = \sigma_{i+1}[Y_i/\delta^{\sigma_{i+1}}(R(Y_i))]$$

Claim. $\sigma_i(Y_i) = \delta^\sigma(pat(G_{Y_i}))$ and σ_i can be computed in time $O(|Q|^3|\sum_{j=i}^n |R(Y_j)|)$ for all $1 \leq i \leq n$.

This claim will imply the Lemma as follows: If $S \in dom(R)$ then $S = Y_i$ for some $1 \leq i \leq n$, so that $\delta^\sigma(pat(G)) = \sigma_i(Y_i)$. Otherwise, $\delta^\sigma(pat(G)) = S$. In both cases, $\delta^\sigma(pat(G))$ can be computed in time $O(|Q|^3|G|)$ by the claim.

We finally prove the claim by a complete induction on i backwards from n to 1.

If $i = n$, then there exists no $Y' \in dom(R)$ such that $Y_n >_G Y'$. This means that $R(Y_n)$ contains no occurrence of an element of $dom(R)$ and by definition, $pat(G_{Y_n}) = R(Y_n)$. On the other hand, $\sigma_n = \sigma_{n+1}[Y_n/\delta^{\sigma_{n+1}}(R(Y_n))]$ which implies that $\sigma_n(Y_n) = \delta^{\sigma_{n+1}}(R(Y_n)) = \delta^\sigma(pat(G_{Y_n}))$. Furthermore, computing σ_n requires $|R(Y_n)|$ multiplications of matrixes, and so $|Q|^3||R(Y_n)|$ steps and time $O(|Q|^3||G|)$.

If $1 \leq i < n$, then $\sigma_i = \sigma_{i+1}[Y_i/\delta^{\sigma_{i+1}}(R(Y_i))]$, and we can write $\sigma_i(Y_i)$ as $\sigma_i(Y_i) = \delta^\sigma(R(Y_i))[Y_j/\sigma_j(Y_j) \mid i < j \leq n]$. By the inductive hypothesis, we have that for all $i < j \leq n$, $\sigma_j(Y_j) = \delta^\sigma(pat(G_{Y_j}))$ and σ_j is computed in time $O(|Q|^3|\sum_{k=j}^n |R(Y_k)|)$. So $\sigma_i(Y_i) = \delta^\sigma(R(Y_i))[Y_j/pat(G_{Y_j}) \mid i < j \leq n] = pat(G_{Y_i})$ and σ_i is computed in $|Q|^3|\sum_{j=i+1}^n |R(Y_j)| + |Q|^3|R(Y_i)|$ steps, that is in time $O(|Q|^3|G|)$ as required.

Proposition 1. INCL$_\Sigma$(*cPat*, NFA) *is in* PSPACE.

Proof. Given an NFA A, a compressed string pattern G over Σ, we have to check whether $Inst(pat(G)) \subseteq \mathcal{L}(A)$. By definition, this is equivalent to checking whether $\delta(\hat{s}(pat(G)))$ is I,F-successful for every substitution $s : fv(G) \to \Sigma^*$. The latter is equivalent to checking whether $\delta^\sigma(pat(G))$ is I,F-successful for all substitution $\sigma : fv(G) \to T_Q$ that maps the free variables of G to δ-inhabited transitions. A decision procedure can thus enumerate all substitutions $\sigma : fv(G) \to T_Q$ to δ-inhabited transitions, compute $\delta^\sigma(pat(G))$ and check whether it is I,F-successful. The number of substitutions σ that is to be checked is exponential, but they can be enumerated in PSPACE. Whether σ maps only to δ-inhabited transition can be tested in PSPACE by Theorem 1. Computing $\delta^\sigma(pat(G))$ can be done in PTIME by Lemma 2.

So far we have shown that $\text{INCL}_\Sigma(cPat, \text{NFA})$ is PSPACE-complete. We next consider regular matching. This will permit us to show that $\text{INCL}_\Sigma(cPat, \text{DFA})$ is PSPACE-hard too.

Proposition 2. $\text{MATCH}_\Sigma(cPat, \text{NFA})$ *is in* PSPACE.

Proof. The proof for $\text{MATCH}_\Sigma(cPat, \text{NFA})$ is similar to that for $\text{INCL}_\Sigma(cPat, \text{NFA})$ in Proposition 1. It is sufficient to check whether there exists some substitution $\sigma : fv(G) \to T_Q$ to δ-inhabited transitions such that $\delta^\sigma(pat(G))$ is I,F-successful.

Proposition 3. $\text{MATCH}_\Sigma(Pat, \text{DFA})$ *is* PSPACE*-hard.*

Proof. This follows by reduction from the non-emptiness problem of the intersection of a sequence of DFA*s*, which is well-known to be PSPACE-complete [17]. Consider a sequence of DFAs $A_1, \ldots, A_n$ with alphabet Σ for some $n \geq 0$ and $\#$ a fresh symbol not in Σ. Let $A = (Q, \Sigma, \delta, I, F)$ be a DFA that recognizes the language $\{u_1\# \ldots \#u_n \mid u_i \in \mathcal{L}(A_i) \text{ for all } 1 \leq i \leq n\}$. Note that such a DFA A can be constructed in linear time from the sequence $A_1 \ldots A_n$. Let p be the pattern $p = y\# \ldots \#y$ with n occurrences of the pattern variable y. We then have $Inst(p) \cap \mathcal{L}(A) \neq \emptyset$ if and only if $\mathcal{L}(A_1) \cap \ldots \cap \mathcal{L}(A_n) \neq \emptyset$.

Lemma 3. *The problem* $\text{coMATCH}_\Sigma(Pat, \text{DFA})$ *is* PSPACE*-complete.*

Proof. Proposition 2 and 3 show the PSPACE-completeness of $\text{MATCH}_\Sigma(\mathcal{G}, \mathcal{A})$ for all $\mathcal{G} \in \{Pat, cPat\}$ and $\mathcal{A} \in \{\text{DFA}, \text{NFA}\}$. Since the complexity class PSPACE is closed by complement, the complemented problems $\text{coMATCH}_\Sigma(\mathcal{G}, \mathcal{A})$ are PSPACE-complete too.

In order to complete the proof of Theorem 2 it remains to relate regular inclusion and matching in the case of DFAs.

Lemma 4. $\text{coMATCH}_\Sigma(\mathcal{G}, \text{DFA}) =_p \text{INCL}_\Sigma(\mathcal{G}, \text{DFA})$ *for all* $\mathcal{G}$ *up to* PTIME *reductions.*

Proof. This follows from $Inst(pat(G)) \cap \mathcal{L}(A) = \emptyset \Leftrightarrow Inst(pat(G)) \subseteq \overline{\mathcal{L}(A)}$ and the fact that any DFA A can be complemented in PTIME to some DFA $\overline{A}$ such that $\overline{L(A)} = L(\overline{A})$.

The complexity of inclusion and matching decreases for linear string patterns, with or without compression. We indeed obtain the same complexity as known for streams (see Fig. 4), even though unknown factors and compression are permitted in addition.

Theorem 3 (Linear patterns). *Restricted to linear string patterns, regular compressed pattern inclusion and matching have the following complexities:*

1. $\text{INCL}_\Sigma(LinPat, \text{NFA})$ *and* $\text{INCL}_\Sigma(cLinPat, \text{NFA})$ *are* PSPACE*-complete if* $|\Sigma| \geq 2$ *while the problem* $\text{INCL}_\Sigma(cLinPat, \text{DFA})$ *can be solved in* PTIME.
2. $\text{MATCH}_\Sigma(cLinPat, \text{NFA})$ *is in* PTIME.

Proof. **1.** The PSPACE-hardness of $\text{INCL}_\Sigma(LinPat, \text{NFA})$ follows from Lemma 1. The problem $\text{INCL}_\Sigma(cLinPat, \text{NFA})$ is in PSPACE by Theorem 2. Hence both problems are PSPACE-complete. By Lemma 4, it is sufficient to demonstrate that $\text{MATCH}_\Sigma(cLinPat, \text{DFA})$ is in PTIME in order to show that $\text{INCL}_\Sigma(cLinPat, \text{DFA})$ is in PTIME, which will follow from point 2 below.
2. We next show that $\text{MATCH}_\Sigma(cLinPat, \text{NFA})$ is in PTIME. In the case without compression this follows from the fact that for any pattern $p \in LinPat_\Sigma$ one can compute in linear time a DFA that recognizes $Inst(p)$. In the case with compression, this argument does not work any more: for compressed string patterns $G \in cLinPat$, the instance set remains regular, but due to compression it may not be possible to represent it by a finite automaton of polynomial size. However, as G is linear, all its string variables can be instantiated independently. So let $A = (Q, \Sigma, \delta, I, F)$ be an NFA. We consider the set of edges of the graph of the transition relation $E_\delta = \cup_{a \in \Sigma} \delta(a)$ and the accessibility relation of this graph $acc = {E_\delta}^*$. Note that acc is a transition in T_Q. Let σ_{acc} be the substitution that maps all string variables in $fv(G)$ to acc. It then holds that $Inst(pat(G)) \cap \mathcal{L}(A) \neq \emptyset$ iff the transition $\delta^{\sigma_{acc}}(pat(G))$ is I,F-successful. This transition can be computed in PTIME by Lemma 2.

5 Defining Queries by Automata

We now recall the notion of queries on strings over some alphabet Σ with *variables* in some finite set $\mathcal{V}$ and relate them to languages of $\mathcal{V}$-annotated strings called $\mathcal{V}$-structures in [26]. We fix two disjoint finite sets Σ and $\mathcal{V}$.

Definition 2 (Query). *A* query *with variables in* $\mathcal{V}$ *on strings over* Σ*, or a* Σ,$\mathcal{V}$-query *for short, is a function* $\boldsymbol{Q}$ *that maps any string* $w \in \Sigma^*$ *to a set* $\boldsymbol{Q}(w)$ *of total assignments from* $\mathcal{V}$ *to* $pos(w)$*. A Boolean query is a* Σ,$\emptyset$*-query.*

Example 3. Let $\mathcal{V} = \{x, x'\}$. The query $\mathbf{Q}_1$ selects all pairs of letters (x, x') such that position x is labeled by a, position x' immediately follows x and is labeled by b. This query then satisfies $\mathbf{Q}_1(aa) = \emptyset$, $\mathbf{Q}_1(ab) = \{[x/1, x'/2]\}$, $\mathbf{Q}_1(abab) = \{[x/1, x'/2], [x/3, x'/4]\}$, etc.

We next show how a Σ,$\mathcal{V}$-query can be identified with a language of $\mathcal{V}$-annotated strings, i.e., of words over the alphabet $\Sigma^{\mathcal{V}} = \Sigma \times 2^{\mathcal{V}}$. A *(query variable) assignment* α to positions of a string $w \in \Sigma^*$ is a partial function from $\mathcal{V}$ to $pos(w)$. We will identify such variable assignments with words whose letters are sets of variables. For any partial function α from $\mathcal{V}$ to $pos(w)$ where $w \in \Sigma^n$, we define a corresponding word in $(2^{\mathcal{V}})^n$ by $word(\alpha) = \alpha^{-1}(1) \dots \alpha^{-1}(n)$. That is, $word(\alpha)[i]$ is the set of variables $x \in dom(\alpha)$ s.t. $\alpha(x) = i$. Furthermore, for any string $w \in \Sigma^*$ and variable assignment α into positions of w, we define the $\mathcal{V}$-annotated string $w{*}\alpha$ as a word over $\Sigma^{\mathcal{V}}$ by $w{*}\alpha = w{*}word(\alpha)$. In examples we will write a^V instead of letters $(a, V) \in \Sigma^{\mathcal{V}}$. For instance, $ab{*}[x/1, x'/2]$ is written as $a^{\{x\}}b^{\{x'\}}$.

Definition 3 ($\mathcal{V}$-structure [26]). *The set of* $\mathcal{V}$-structures *over* Σ *is the following set of* $\mathcal{V}$*-annotated strings, i.e., of words over* $\Sigma^{\mathcal{V}}$*:*

$$Struct^{\mathcal{V}} = \{w{*}\alpha \in (\Sigma^{\mathcal{V}})^* \mid w \in \Sigma^*,\ \alpha : \mathcal{V} \to pos(w)\}.$$

We note that all the assignments α in the definition of $\mathcal{V}$-structures are total functions. For instance for $\mathcal{V} = \{x, x'\}$ and $\Sigma = \{a, b\}$, the words $a^{\emptyset}b^{\{x',x\}}$ and $a^{\{x'\}}b^{\{x\}}$ are $\mathcal{V}$-structures while the words $a^{\emptyset}b^{\{x'\}}$ and $a^{\{x'\}}b^{\{x',x\}}$ are not. Essentially, $\mathcal{V}$-structures represent total variable assignments to the positions of a string without naming the positions.

Definition 4 (Language of $\mathcal{V}$-structures of a query). *For any* Σ,$\mathcal{V}$*-query* $\boldsymbol{Q}$*, the* language of $\mathcal{V}$-structures of $\boldsymbol{Q}$ *is* $\mathcal{L}(\boldsymbol{Q}) = \{w{*}\alpha \mid w \in \Sigma^*,\ \alpha \in \boldsymbol{Q}(w)\}$.

We will be interested in queries whose languages are definable by NFAs.

Definition 5 (Query automata). *An* Σ,$\mathcal{V}$*-query automaton is an* NFA A *such that* $\mathcal{L}(A)$ *is a language of* $\mathcal{V}$*-structures over* Σ*. The unique* Σ,$\mathcal{V}$*-query such that* $\mathcal{L}(\boldsymbol{Q}) = \mathcal{L}(A)$ *is called the query defined by* A *and is denoted by* $\mathcal{Q}(A) = \boldsymbol{Q}$*.*

6 Certain Query Answers and Non-Answers

We next formalize the notions of certain query answers and non-answers on string patterns. For streams, these definitions coincide with the notions of earliest query answers from [12] and fast-failure from [4], respectively.

A Σ*-assignment* for $\mathcal{V}$ on $p \in Pat_{\Sigma}$ is a partial function α from $\mathcal{V}$ to $pos_{\Sigma}(p)$, the Σ-positions of the pattern p. For any Σ-assignment α on p, the word $p{*}\alpha$ is a string pattern over $\Sigma^{\mathcal{V}}$, still with string variables in $\mathcal{Y}$. Therefore, the set $Inst(p{*}\alpha)$ is a well-defined set of words over $\Sigma^{\mathcal{V}}$. Note, however, that some of these $\mathcal{V}$-annotated strings may not be $\mathcal{V}$-structures. For instance, if $x \in \mathcal{V}$, $a \in \Sigma$, and $Y \in \mathcal{Y}$, then $a^{\{x\}}a^{\{x\}} \in Inst(Y{*}[])$ is not a $\mathcal{V}$-structure since x occurs twice (where $[]$ is the empty Σ-assignment).

Definition 6 (Certain query answers and non-answers). *Let* $\boldsymbol{Q}$ *be a* Σ,$\mathcal{V}$*-query, and let* $p \in Pat_{\Sigma}$ *be a string pattern. A* Σ*-assignment* α *for* $\mathcal{V}$ *on* p *is:*

- *a* certain answer *for query* $\boldsymbol{Q}$ *on* p *if* α *is total and* $Inst(p{*}\alpha) \cap Struct^{\mathcal{V}} \subseteq \mathcal{L}(\boldsymbol{Q})$,
- *and a* certain non-answer *for query* $\boldsymbol{Q}$ *on* p *if* $Inst(p{*}\alpha) \cap \mathcal{L}(\boldsymbol{Q}) = \emptyset$.

Given an instance $w \in Inst(p)$, each Σ-assignment α on p defines a set of total Σ-assignments on w, where all variables not in $dom(\alpha)$ must be mapped to some Σ-positions "created" by the instantiation. More formally:

$$C_{p,w}(\alpha) = \{\alpha' \mid \alpha' \text{ is a total } \Sigma\text{-assignment on } w,\ w{*}\alpha' \in Inst(p{*}\alpha)\}.$$

The offsets of positions of query variables due to the instantiation of pattern variables raise two issues that we illustrate in the following example: 1. even a total Σ-assignment α of p might have several completions for the same string w, and 2. it might be the case that $\alpha \notin C_{p,w}(\alpha)$.

Example 4. Consider the string pattern $p = ay_1ay_2$, the string $w = aaba$ in $Inst(p)$, $\mathcal{V} = \{x\}$, and the total Σ-assignment $\alpha = [x/3]$ on p. In order to make p match w, the second a-letter of p matches either the second or the fourth position in w. Therefore, there are exactly two substitutions that make p match w, which are $\sigma_1 = [y_1/\varepsilon, y_2/ba]$ and $\sigma_2 = [y_1/ab, y_2/\varepsilon]$. Now, $\hat{\sigma_1}(p{*}\alpha) = \hat{\sigma_1}(ay_1a^{\{x\}}y_2) = aa^{\{x\}}ba = w{*}[x/2]$, thus $[x/2] \in C_{p,w}(\alpha)$. Also, $\hat{\sigma_2}(p{*}\alpha) = \hat{\sigma_1}(ay_1a^{\{x\}}y_2) = aaba^{\{x\}} = w{*}[x/4]$, thus $[x/4] \in C_{p,w}(\alpha)$. Given that there is no further way to match p with w, there is no further completion. That is, $C_{p,w}(\alpha) = \{[x/2], [x/4]\}$, and $\alpha \notin C_{p,w}(\alpha)$.

The next proposition relates certain query answers (resp. non-answers) on a pattern p to query answers (resp. non-answers) on its instances.

Proposition 4. *Let* α *be a* Σ*-assignment on string pattern* p *and* $\boldsymbol{Q}$ *be a* Σ,$\mathcal{V}$*-query. It then holds for all instances* $w \in Inst(p)$*:*

- *If* α *is a certain answer for query* $\boldsymbol{Q}$ *on* p *then* $C_{p,w}(\alpha) \subseteq \boldsymbol{Q}(w)$.
- *If* α *is a certain non-answer for query* $\boldsymbol{Q}$ *on* p *then* $C_{p,w}(\alpha) \cap \boldsymbol{Q}(w) = \emptyset$.

Proof. For all $\alpha' \in C_{p,w}(\alpha)$, the definition of completion yields that $w{*}\alpha' \in Inst(p{*}\alpha)$. Furthermore, $w{*}\alpha' \in Struct^{\mathcal{V}}$ holds trivially, since it holds for any partial function α'.

- If α is a certain answer for query $\mathbf{Q}$ on p then $Inst(p{*}\alpha) \cap Struct^{\mathcal{V}} \subseteq \mathcal{L}(Q)$, so $\alpha' \in \mathcal{L}(\mathbf{Q})$ which is equivalent to $\alpha' \in \mathbf{Q}(w)$.
- If α is a certain non-answer for query $\mathbf{Q}$ then $Inst(p{*}\alpha) \cap \mathcal{L}(Q) = \emptyset$, so $w{*}\alpha' \notin \mathcal{L}(Q)$, which is equivalent to $\alpha' \notin \mathbf{Q}(w)$.

7 Certain Query Answering and Non-Answering

We introduce the problems of certain query answering and non-answering for classes of compressed string patterns $\mathcal{G}$ and of query NFAs $\mathcal{A}$:

Certain query answering $\textsc{Cert}^{ans}_{\Sigma,\mathcal{V}}(\mathcal{G},\mathcal{A})$. *Input:* a compressed string pattern $G \in \mathcal{G}_\Sigma$, a Σ-assignment α for $\mathcal{V}$ on $pat(G)$, and a query Nfa $A \in \mathcal{A}_{\Sigma^\mathcal{V}}$.
Output: whether α is a certain answer for query $\mathcal{Q}(A)$ on $pat(G)$.

Certain query non-answering $\textsc{Cert}^{\neg ans}_{\Sigma,\mathcal{V}}(\mathcal{G},\mathcal{A})$. *Input:* as above
Output: whether α is a certain non-answer for query $\mathcal{Q}(A)$ on $pat(G)$.

Note that α is an assignment to positions of $pat(G)$ and not to positions of G. This is necessary because due to compression, a position of G might correspond to several positions of the underlying word which need to be distinguished. Indeed, considering G to be the compressed string pattern in the introduction, the a-letter in the rule for X corresponds to positions 5 and 13 in the decompressed pattern $a\,YcZa\,bb\,Ya\,YcZa$, and position 5 is a certain answer for query $\mathbf{Q}_1$, while position 13 is a certain non-answer.

The following lemma relates certain query answering to regular inclusion and certain query non-answering to regular matching.

Lemma 5 (Boolean Queries). *For any classes $\mathcal{G}$ and $\mathcal{A}$,* $\textsc{Incl}_\Sigma(\mathcal{G},\mathcal{A}) =_p \textsc{Cert}^{ans}_{\Sigma,\emptyset}(\mathcal{G},\mathcal{A})$ *and* $\textsc{coMatch}_\Sigma(\mathcal{G},\mathcal{A}) =_p \textsc{Cert}^{\neg ans}_{\Sigma,\emptyset}(\mathcal{G},\mathcal{A})$.

Proof. Straightforward from the definitions.

Since PSpace-complete problems are closed by complement, Lemma 5 implies together with Theorem 2 that $\textsc{Cert}^{ans}_{\Sigma,\emptyset}(Pat,\textsc{Dfa})$ and $\textsc{Cert}^{\neg ans}_{\Sigma,\emptyset}(Pat,\textsc{Dfa})$ are PSpace-complete, even though these problems are restricted to Dfas, without compression, and for Boolean queries. Therefore, all certainty problems $\textsc{Cert}^{\mathcal{B}}_{\Sigma,\emptyset}(\mathcal{G},\mathcal{A})$ where $\mathcal{B} \in \{\neg ans, ans\}$, $\mathcal{G} \in \{Pat, cPat\}$, and $\mathcal{A} \in \{\textsc{Dfa},\textsc{Nfa}\}$ are PSpace-hard. In the sequel we show that all these problems can be solved in PSpace for arbitrary finite sets Σ and $\mathcal{V}$. Basically, these results will be corollaries of Theorem 2, Lemma 5 on Boolean queries, and the following partial decompression lemma. This result is new to the best of our knowledge, even though its proof relies on similar techniques as used for instance in [5] for computing in PTime the letter at the n-th position of $pat(G)$ for a singleton grammar G.

Lemma 6 (Partial Decompression). *For any $G \in cPat_\Sigma$ and any Σ-assignment α for $\mathcal{V}$ on $pat(G)$, we can compute in* PTime *some $G' \in cPat_{\Sigma^\mathcal{V}}$ such that $pat(G){*}\alpha = pat(G')$. In particular, if $pat(G)$ was linear then $pat(G')$ is linear.*

Let Σ and $\mathcal{V}$ be finite sets, $G = (N, \Sigma, R, S)$ be a compressed string pattern and α be a Σ-assignment for $\mathcal{V}$ on $pat(G)$, fixed in this section. We show that we can compute in PTime a compressed string pattern G' over $\Sigma^\mathcal{V}$ such that $pat(G){*}\alpha = pat(G')$. We assume that $S \in dom(R)$ as otherwise the lemma is trivial.

The main ingredients of the proof will be illustrated on the example compressed string pattern $G = (\{S, Y_1, Y_2\}, \{a, b\}, R, S)$ in Fig. 7, where $R(S) = aY_1aY_1Y_1$ and $R(Y_1) = bY_2$. We also use for the example $\mathcal{V} = \{x\}$ and $\alpha = [x/5]$. Note that $pat(G) = abY_2abY_2bY_2$ and $pat(G){*}\alpha = a^\emptyset b^\emptyset Y_2 a^\emptyset b^{\{x\}} Y_2 b^\emptyset Y_2$.

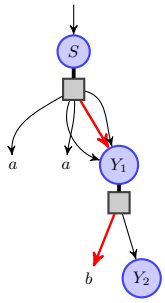

Fig. 7: G with red thick path to the shared position 5.

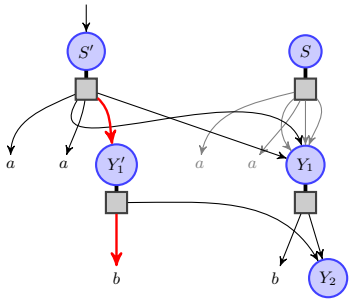

Fig. 8: Partially uncompressing G at position 5.

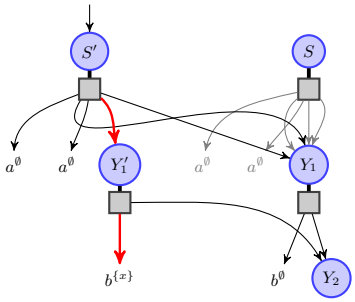

Fig. 9: G' on $\Sigma^{\mathcal{V}}$ with $pat(G') = pat(G)*[x/5]$.

First we define the *addresses* of G that are non-empty words over the alphabet $\mathbb{N}$ of natural numbers, defined similarly to the standard Dewey notation for trees but applied to the acyclic graph structure of G. We define $Addr = Addr(S)$ where for any $Y \in dom(R)$, denoting $s = R(Y)$:

$$Addr(Y) = pos(s) \cup \left\{ (Y,i) \cdot d \mid i \in pos_{dom(R)}(s) \text{ and } d \in Addr(s[i]) \right\}.$$

On the example, $d_1 = 22$, $d_2 = 41$ and $d_3 = 51$ are addresses. With any address d we associate its path that is a word whose letters are pairs of the form (Y,i) for $Y \in dom(R)$ and $i \in pos(R(Y))$: $path(k_1 \dots k_n) = (Y_1,k_1),\dots,(Y_n,k_n)$ where $Y_1 = S$ and $Y_i = R(Y_{i-1})[k_i]$ for any $2 \le i \le n$. On the example, $path(d_1) = (S,2)(Y_1,2)$, $path(d_2) = (S,4)(Y_1,1)$ (the thick red arrows) and $path(d_3) = (S,5)(Y_1,1)$. Note that $path$ is injective, therefore $path^{-1}$ is defined. Let $Addr_\Sigma$ be the set of addresses of G that lead to a letter in Σ. That is, an address $d \cdot (Y,i)$ is in $Addr_\Sigma$ if $R(Y)[i] \in \Sigma$.

We first establish that there is a one-to-one correspondence between $Addr_\Sigma$ and $pos_\Sigma(pat(G))$, the Σ-positions of $pat(G)$.

Claim. There is a bijection $addr : pos_\Sigma(pat(G)) \to Addr_\Sigma$ s.t. for any $m \in pos_\Sigma(pat(G))$, $addr(m)$ can be computed in PTIME in the size of G.

Proof. For any $U \in \Sigma \cup N$, we define $patsize(U)$ that is, intuitively, the size of the pattern produced by G starting at U. Formally, $patsize(a) = 1$ for any $a \in \Sigma$, $patsize(Y) = 0$ for any $Y \in fv(G)$, and $patsize(Y) = |pat(G_U)|$ for any $U \in dom(R)$. Now for any (Y, i) s.t. $Y \in dom(R)$ and $i \in pos(R(Y))$, we define $offset(Y, i) = \sum_{j<i} patsize(R(Y)[j])$, that intuitively is the size of the pattern of G_Y if $R(Y)$ was truncated just before its i^{th} position. Finally, for any $d \in Addr$, we define $offset(d) = \sum_{i \in pos(d)} offset(d[i])$.

Now we define and show how to compute $addr'$ that with any $m \in pos(pat(G))$ and $Y \in dom(R)$ associates a maximal address among the addresses ofG_Y. An address is maximal if it is not the prefix of any other address. First, for any (Y, i) s.t. $i \in pos(R(Y))$ we compute and memorize $offset(Y, i)$. This can be done in time $O(|G|)$ using the acyclic relation $>_G$. Then $addr'(m, Y)$ is computed recursively by descending the tree structure induced by the addresses of G_Y:

- $addr'(m, Y) = (Y, i)$ if $dom(R) \cap fv(R(Y)) = \emptyset$, where i is the least value among $pos(R(Y))$ s.t. $offset(Y, i) > m$;
- $addr'(m, Y) = (Y, i) \cdot addr'(m', Y')$ otherwise, where i is as in the previous case and $m' = m - offset(Y, i)$.

We define $addr(m) = addr'(m, S)$ for any $m \in pos_\Sigma(pat(G))$. From the definition it is easy to see that $addr(m)$ can be computed in PTIME in the size of G. We claim that $addr(m)$ is a bijection such that $offset(addr(m)) = m - 1$.

On the example, $addr(5) = d_2$ and $addr(7) = d_3$.

The next ingredient for the proof of the partial decompression lemma is to show that given some address, G can be transformed into G' having the same pattern, but in which this address is sharing free. An address $d \cdot U \in Addr$ is called *sharing-free* if there does not exist a different address $d' \cdot U \in Addr$ s.t. $path(d)$ and $path(d')$ have the same symbol in their last position.

Claim. For any address $d \in NumAddr$ we can compute in PTIME in the sizes of G and d a compressed string pattern G'' such that $pat(G'') = pat(G)$ and d is sharing-free in G''.

Proof. Let $\mathcal{Y}_d \subseteq \mathcal{Y}$ be the set of variables $\{Y_1, \ldots, Y_n\}$ such that $path(d) = (Y_1, k_1) \cdots (Y_n, k_n)$ for some naturals $k_1, \ldots, k_n$. Let $\mathcal{Y}''$ be a set with same cardinality as $\mathcal{Y}_d$ and disjoint from N, and let $copy : \mathcal{Y}_d \to \mathcal{Y}''$ be a bijection. We define the compressed string pattern $G'' = (N \cup \mathcal{Y}'', \Sigma, R'', copy(S))$ with $dom(R'') = dom(R) \cup \mathcal{Y}''$ and:

- if $Y \in dom(R)$, then $R''(Y) = R(Y)$,
- otherwise, let $s = R(copy^{-1}(Y)) = (s_1, \ldots, s_m)$, i.e. s is the word over $\Sigma \cup N$ that is the definition of $copy^{-1}(Y)$ in G. Then $R''(Y) =_{def} (s'_1, \ldots, s'_m)$ is of same length as s and its i^{th} position is $s'_i = copy(s_i)$ if $s_i \in \mathcal{Y}_d$ and (Y, i) is a letter in $path(d)$, and $s'_i = s_i$ otherwise.

Then it is not hard to see that $path(d)$ in the compressed string pattern G'' is equal to $(copy(Y_1), k_1) \cdots (copy(Y_n), k_n)$. Thus d is sharing-free in G'' as any variable $copy(Y)$ for $Y_i \in \mathcal{Y}_d$ for $2 \leq i \leq n$ is used only once in R'' in $R''(Y_{i-1})$. It is also not hard to show that for any $Y \in \mathcal{Y}_d$, $pat(G''_{copy(Y)}) = pat(G''_Y) = pat(G_Y)$, thus $pat(G) = pat(G'')$.

Reconsider Fig. 7. Partial decompression of G for unsharing position 5 at address $d_2 = 41$ (the red path) yields the compressed string pattern in Fig. 8. Finally, as last ingredient of the proof of the partial decompression lemma, we show how for a given Σ-assignment α, a compressed string pattern Σ is transformed to a compressed string pattern on $\Sigma^{\mathcal{V}}$ in which the positions associated with query variables by α are sharing-free.

Using the bijection $addr : pos_\Sigma(pat(G)) \to Addr$ defined in Claim 7:

Claim. If all addresses in $addr(\alpha(\mathcal{V}))$ are sharing-free in G, we can compute in PTIME a compressed string pattern G' over $\Sigma^{\mathcal{V}}$ such that $pat(G') = pat(G)*\alpha$.

Proof. For all $m \in pos_\Sigma(G)$, let $last(addr(m))$ be the symbol from $N \times \mathbb{N}$ at the last position in $path(addr(m))$. We define $G' = (N, \Sigma^{\mathcal{V}}, R', S)$ by $dom(R') = dom(R)$, and for any Y in $dom(R)$, $R'(Y)$ has the same length as $R(Y)$ and for any $j \in pos(R(Y))$ we set

$$R'(Y)[j] = \begin{cases} a^{\mathcal{S}} & \text{if } R(Y)[j] = a \in \Sigma, \\ & \text{where } \mathcal{S} = \{x \in dom(\alpha) \mid last(addr(\alpha(x))) = (Y, j)\} \\ R(Y) & \text{otherwise} \end{cases}$$

Since all positions in $addr(\alpha(\mathcal{V}))$, we have that $pat(G')$ is a correct $\mathcal{V}$-structure. It is also tedious but not hard to prove that $pat(G') = pat(G)*\alpha$.

On the example, the compressed string pattern in Figure 8 satisfies the hypothesis of the latter claim, and its corresponding pattern on G' on $\Sigma^{\mathcal{V}}$ is depicted in Figure 9.

Proof of Lemma 6. We first compute the set of addresses $\mathcal{D} = addr(\alpha(\mathcal{V}))$ in PTIME by Claim 7. Note that the cardinality of $\mathcal{D}$ is at most $|\mathcal{V}|$, so it is of constant size. Then we compute a compressed string pattern G'' with $pat(G) = pat(G'')$ such that all addresses in $num(\mathcal{D})$ are sharing-free in G''. Since there are constantly many such addresses, this can be done in PTIME by iterating Claim 7 a constant number of times. Then using Claim 7 we compute the compressed string pattern G' over $\Sigma^{\mathcal{V}}$ such that $pat(G') = pat(G'')*\alpha = pat(G)*\alpha$.

Proposition 5. *For all $\mathcal{B}$ in $\{ans, \neg ans\}$, all $\mathcal{G}$ in $\{Pat, cPat, LinPat, cLinPat\}$, and all $\mathcal{A}$ in $\{\text{DFA}, \text{NFA}\}$, there is a* PTIME *reduction from* $\text{CERT}^{\mathcal{B}}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A})$ *to* $\text{CERT}^{\mathcal{B}}_{\Sigma^{\mathcal{V}},\emptyset}(\mathcal{G}, \mathcal{A})$.

Proof. Let $G \in \mathcal{G}_\Sigma$ be a compressed string pattern and α a Σ-assignment from $\mathcal{V}$ to positions of $pat(G)$. If $\mathcal{G} \in \{cPat, cLinPat\}$, then by Lemma 6 we can compute in PTIME a compressed string pattern $G' \in \mathcal{G}_{\Sigma^{\mathcal{V}}}$ such that $pat(G)*\alpha = pat(G')$. If $\mathcal{G} \in \{Pat, LinPat\}$, then the same property holds trivially.

We start with $\mathcal{B} = ans$. Let $A \in \mathcal{A}_{\Sigma^\mathcal{V}}$ be a query automaton, $B \in \text{DFA}_{\Sigma^\mathcal{V}}$ such that $L(B) = \Sigma^{\mathcal{V}^*} \setminus \mathit{Struct}^\mathcal{V}$, and $C \in \mathcal{A}_{\Sigma^\mathcal{V}}$ such that $L(C) = L(A) \cup L(B)$. In the case of $\mathcal{G} = \text{NFA}$ we can chose C to be the union of A and B and in the case of $\mathcal{G} = \text{DFA}$, C can be chosen as the product of A and B. The automaton B can be constructed in time time $O(2^{|\mathcal{V}|})$ which is constant since $\mathcal{V}$ is a parameter of the certainty problem rather than being part of its input. Then α is a certain answer of $\mathcal{Q}(A)$ on G iff $\mathit{Inst}(G{*}\alpha) \cap \mathit{Struct}^\mathcal{V} \subseteq \mathcal{L}(A)$ iff $\mathit{Inst}(G') \cap \mathit{Struct}^\mathcal{V} \subseteq \mathcal{L}(A)$ iff $\mathit{Inst}(G') \subseteq \mathcal{L}(A) \cup L(B)$ iff $\mathit{Inst}(G') \subseteq \mathcal{L}(C)$ iff the empty Σ-assignment is a certain answer of $\mathcal{Q}(C)$ on G'. Based on this fact, the PTIME reduction $\text{CERT}^{ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A}) \leq_p \text{CERT}^{ans}_{\Sigma^\mathcal{V},\emptyset}(\mathcal{G}, \mathcal{A})$ is obvious. Also, α is a certain non-answer of $\mathcal{Q}(A)$ on G iff $\mathit{Inst}(G{*}\alpha) \cap \mathcal{L}(A) = \emptyset$ iff $\mathit{Inst}(G') \cap \mathcal{L}(A) = \emptyset$ iff the empty Σ-assignment is a certain non-answer of $\mathcal{Q}(A)$. This yields the PTIME reduction showing $\text{CERT}^{\neg ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A}) \leq_p \text{CERT}^{\neg ans}_{\Sigma^\mathcal{V},\emptyset}(\mathcal{G}, \mathcal{A})$.

Corollary 1 (Non-linear patterns). *For all $\mathcal{B} \in \{ans, \neg ans\}$, $\mathcal{G} \in \{Pat, cPat\}$ and $\mathcal{A} \in \{\text{DFA}, \text{NFA}\}$, the problem $\text{CERT}^{\mathcal{B}}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A})$ is* PSPACE*-complete.*

Proof. Let $\mathcal{A} \in \{\text{DFA}, \text{NFA}\}$ and $\mathcal{G} \in \{Pat, cPat\}$. We have that $\text{INCL}_\Sigma(\mathcal{G}, \mathcal{A}) =_p \text{CERT}^{ans}_{\Sigma,\emptyset}(\mathcal{G}, \mathcal{A}) \subseteq \text{CERT}^{ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A}) \leq_p \text{CERT}^{ans}_{\Sigma^\mathcal{V},\emptyset}(\mathcal{G}, \mathcal{A}) \leq_p \text{INCL}_{\Sigma^\mathcal{V}}(\mathcal{G}, \mathcal{A})$. Since $\text{INCL}_\Sigma(\mathcal{G}, \mathcal{A})$ and $\text{INCL}_{\Sigma^\mathcal{V}}(\mathcal{G}, \mathcal{A})$ are PSPACE-complete by Theorem 2, the problem $\text{CERT}^{ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A})$ is also PSPACE-complete. On the other hand, we have that $\text{coMATCH}_\Sigma(\mathcal{G}, \mathcal{A}) =_p \text{CERT}^{\neg ans}_{\Sigma,\emptyset}(\mathcal{G}, \mathcal{A}) \subseteq \text{CERT}^{\neg ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A}) \leq_p \text{coMATCH}_{\Sigma^\mathcal{V}}(\mathcal{G}, \mathcal{A})$ by Lemma 5 and Proposition 5. Then by Theorem 2, $\text{coMATCH}_\Sigma(\mathcal{G}, \mathcal{A})$ and $\text{coMATCH}_{\Sigma^\mathcal{V}}(\mathcal{G}, \mathcal{A})$ are PSPACE-complete (since PSPACE is closed by complementation), and thus $\text{CERT}^{\neg ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \mathcal{A})$ is PSPACE-complete. This completes the proof.

Our next objective is to study the complexity of the four certainty problems but restricted to compressed linear string patterns.

Corollary 2 (Linear patterns). *For all $\mathcal{A} \in \{\text{DFA}, \text{NFA}\}$, the certainty problems $\text{CERT}^{\neg ans}_{\Sigma,\mathcal{V}}(cLinPat, \mathcal{A})$ (1) and $\text{CERT}^{ans}_{\Sigma,\mathcal{V}}(cLinPat, \text{DFA})$ (2) can be solved in* PTIME*. The problem $\text{CERT}^{ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \text{NFA})$ is* PSPACE*-complete for all $\mathcal{G} \in \{LinPat, cLinPat\}$ (3).*

Proof. The statements are proved respectively in (1), (2) and (3):

(1) For all $\mathcal{A} \in \{\text{DFA}, \text{NFA}\}$, we have that $\text{CERT}^{\neg ans}_{\Sigma,\mathcal{V}}(cLinPat, \mathcal{A}) \leq_p \text{CERT}^{\neg ans}_{\Sigma^\mathcal{V},\emptyset}(cLinPat, \mathcal{A}) \leq_p \text{coMATCH}_{\Sigma^\mathcal{V}}(cLinPat, \mathcal{A}) \in$ PTIME, so $\text{CERT}^{\neg ans}_{\Sigma,\mathcal{V}}(cLinPat, \mathcal{A}) \in$ PTIME.

(2) follows Theorem 3 and the reductions $\text{CERT}^{ans}_{\Sigma,\mathcal{V}}(cLinPat, \text{DFA}) \leq_p \text{CERT}^{ans}_{\Sigma^\mathcal{V},\emptyset}(cLinPat, \text{DFA}) \leq_p \text{INCL}_{\Sigma^\mathcal{V}}(cLinPat, \text{NFA}) \in$ PTIME.

(3) For all $\mathcal{G} \in \{LinPat, cLinPat\}$, $\text{INCL}_\Sigma(\mathcal{G}, \text{NFA}) =_p \text{CERT}^{ans}_{\Sigma,\emptyset}(\mathcal{G}, \text{NFA}) \subseteq \text{CERT}^{ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \text{NFA}) \leq_p \text{CERT}^{ans}_{\Sigma^\mathcal{V},\emptyset}(\mathcal{G}, \text{NFA}) \leq_p \text{INCL}_{\Sigma^\mathcal{V}}(\mathcal{G}, \text{NFA})$. Since for all $\mathcal{G} \in \{LinPat, cLinPat\}$, $\text{INCL}_\Sigma(\mathcal{G}, \text{NFA})$ and $\text{INCL}_{\Sigma^\mathcal{V}}(\mathcal{G}, \text{NFA})$ are PSPACE-complete by Theorem 3, we have that $\text{CERT}^{ans}_{\Sigma,\mathcal{V}}(\mathcal{G}, \text{NFA})$ is PSPACE-complete.

8 Conclusion

There exist highly efficient streaming algorithms for answering queries defined by NFAs [10] on complex event streams with low latency, but not with lowest latency, since they approximate the sets of certain query answers at any event. The positive results presented here yield good hope that similar algorithms could be developed for hyperstreams when approximated by compressed linear string patterns. As shown by the authors in a follow-up paper, the linearity restriction is not sufficient. But with a further restriction it is possible to approximate certain query answers efficiently and with decent precision [8]. However, this still requires more research. First, one needs to understand how such algorithms may deal with unknown factors incrementally, without requiring cubic time per step such as previous algorithms on incremental evaluation of queries defined by NFAs [6]. Second, one has to understand how to deal with nested word automata rather than NFAs for dealing with regular path queries on complex event streams. Another point is to develop streaming algorithms for hyperstreams with data values from an infinite signature. Finally, the feasibility of hyperstreaming algorithms needs to be proven in practice.

References

1. R. Angles, M. Arenas, P. Barceló, A. Hogan, J. L. Reutter, and D. Vrgoc. Foundations of modern graph query languages. *CoRR*, abs/1610.06264, 2016.
2. D. Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62, 1980.
3. L. Babai and E. Szemeredi. On the complexity of matrix group problems i. In *Proceedings of the 25th Annual Symposium onFoundations of Computer Science, 1984*, SFCS '84, pages 229–240, Washington, DC, USA, 1984. IEEE Computer Society.
4. M. Benedikt, A. Jeffrey, and R. Ley-Wild. Stream Firewalling of XML Constraints. In *ACM SIGMOD International Conference on Management of Data*, pages 487–498. ACM-Press, 2008.
5. P. Bille, G. M. Landau, R. Raman, K. Sadakane, S. R. Satti, and O. Weimann. Random access to grammar-compressed strings and trees. *SIAM J. Comput.*, 44(3):513–539, 2015.
6. H. Björklund, W. Gelade, and W. Martens. Incremental XPath evaluation. *ACM Trans. Database Syst.*, 35(4):29, 2010.
7. M. Blondin, A. Krebs, and P. McKenzie. The complexity of intersecting finite automata having few final states. *computational complexity*, 25(4):775–814, Dec 2016.
8. I. Boneva, J. Niehren, and M. Sakho. Approximating Certain Query Answering on Hyperstreams, June 2018. technical report.
9. C. David, L. Libkin, and F. Murlak. Certain answers for XML queries. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, June 6-11, 2010, Indianapolis, Indiana, USA*, pages 191–202. ACM, 2010.

10. D. Debarbieux, O. Gauwin, J. Niehren, T. Sebastian, and M. Zergaoui. Early nested word automata for xpath query answering on XML streams. *Theor. Comput. Sci.*, 578:100–125, 2015.
11. A. Gascón, G. Godoy, and M. Schmidt-Schauß. Context matching for compressed terms. In *Proceedings of the Twenty-Third Annual IEEE Symposium on Logic in Computer Science, LICS 2008, 24-27 June 2008, Pittsburgh, PA, USA*, pages 93–102. IEEE Computer Society, 2008.
12. O. Gauwin and J. Niehren. Streamable fragments of forward XPath. In B. B. Markhoff, P. Caron, J. M. Champarnaud, and D. Maurel, editors, *International Conference on Implementation and Application of Automata*, volume 6807 of *Lecture Notes in Computer Science*, pages 3–15. Springer, 2011.
13. O. Gauwin, J. Niehren, and S. Tison. Earliest Query Answering for Deterministic Nested Word Automata. In *17th International Symposium on Fundamentals of Computer Theory*, volume 5699 of *LNCS*, pages 121–132, Wraclaw, Poland, Sept. 2009. Springer Verlag.
14. O. Gauwin, J. Niehren, and S. Tison. Queries on XML streams with bounded delay and concurrency. *Information and Computation*, 209:409–442, 2011.
15. T. J. Green, A. Gupta, G. Miklau, M. Onizuka, and D. Suciu. Processing XML streams with deterministic automata and stream indexes. *ACM Trans. Database Syst.*, 29(4):752–788, Dec. 2004.
16. M. Kay. A streaming XSLT processor. In *Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies*, volume 5, 2010.
17. D. Kozen. Lower bounds for natural proof systems. In *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977*, pages 254–266. IEEE Computer Society, 1977.
18. V. Kumar, P. Madhusudan, and M. Viswanathan. Visibly pushdown automata for streaming XML. In *16th international conference on World Wide Web*, pages 1053–1062. ACM-Press, 2007.
19. O. Kupferman and M. Y. Vardi. Model checking of safety properties. *Formal Methods in System Design*, 19(3):291–314, 2001.
20. P. Labath and J. Niehren. A functional language for hyperstreaming XSLT. Technical report, INRIA Lille, 2013.
21. S. Maneth, A. O. Pereira, and H. Seidl. Transforming XML streams with references. In C. S. Iliopoulos, S. J. Puglisi, and E. Yilmaz, editors, *String Processing and Information Retrieval - 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings*, volume 9309 of *Lecture Notes in Computer Science*, pages 33–45. Springer, 2015.
22. B. Mozafari, K. Zeng, and C. Zaniolo. High-performance complex event processing over XML streams. In K. S. Candan, Y. Chen, R. T. Snodgrass, L. Gravano, A. Fuxman, K. S. Candan, Y. Chen, R. T. Snodgrass, L. Gravano, and A. Fuxman, editors, *SIGMOD Conference*, pages 253–264. ACM, 2012.
23. D. Olteanu. SPEX: Streamed and progressive evaluation of XPath. *IEEE Trans. on Know. Data Eng.*, 19(7):934–949, 2007.
24. W. Plandowski. *The complexity of the morphism equivalence problem for context-free languages.* PhD thesis, Warsaw University. Department of Informatics, Mathematics, and Mechanics, 1995.
25. M. Schmidt, S. Scherzinger, and C. Koch. Combined static and dynamic analysis for effective buffer minimization in streaming XQuery evaluation. In *23rd IEEE International Conference on Data Engineering*, pages 236–245, 2007.
26. H. Straubing. *Finite Automata, Formal Logic, and Circuit Complexity.* Progress in Computer Science and Applied Series. Birkhäuser, 1994.

Acknowledgments. We are thankful to C. Paperman, who saw the PSPACE-hardness of regular string pattern matching in a discussion on the topic. We thank S. Salvati and S. Tison for discussions on regular string pattern matching. It is a pleasure to thank all the anonymous reviewers for their extraordinary helpful feedback.